DOI：10.13998/j.cnki.issn1002-1248.23-0757

期刊下载因子：传播、影响、知识与信息量的综合指标

——提高文献计量指标时效性新的尝试

俞立平[1,2]

（1. 浙江工商大学 统计与数学学院，杭州 310018；2. 浙江工商大学 统计数据工程技术与应用协同创新中心，杭州 310018）

摘　要：[目的 / 意义]综合表征学术期刊传播、影响、知识与信息量的指标缺乏，本文提出下载因子指标以弥补这个问题。[方法 / 过程]首先根据下载频次与被引频次的历年变化，基于图书馆情报与文献学 CSSCI 期刊中国知网引文数据，采用面板数据模型建立下载频次与被引频次预测模型，确定设计下载因子的最佳滞后期，提出下载因子指标，即期刊论文发表 2 年后平均每篇论文的下载次数除以 100。继续采用岭回归分析下载因子与影响因子、h 指数、载文量的关系。[结果 / 结论]滞后 1 年和 2 年下载频次决定了被引频次的 80%；下载因子可以较好测度期刊的知识信息量、传播水平、影响力和学术质量；下载因子指标有待更多学科和数据的检验。

关键词：下载因子；传播水平；滞后期；评价指标

中图分类号：G302　　　　文献标识码：A　　　　文章编号：1002-1248（2023）11-0077-09

引用本文：俞立平. 期刊下载因子：传播、影响、知识与信息量的综合指标——提高文献计量指标时效性新的尝试[J]. 农业图书情报学报, 2023, 35(11)：77-85.

1 引 言

互联网的诞生给文献计量学带来了革命性的影响，催生了一批学术文献的网络下载指标。其中最具有代表性的基础指标就是下载频次，此外还包括 Web 即年下载率、总下载量、下载量半衰期、谷歌学者指数等，这些指标的提出提供了一种新的衡量学术传播和影响力的方法与手段，大大发展了传统的文献计量学，并且也是替代计量学（Altmetrics）的重要组成部分。

目前关于下载频次的相关指标研究还不充分。尽管 GARFIELD[1]早在 1996 年就提出了利用下载频次代替被引频次指标以解决引文分析评价中的滞后性问题的思想，但这方面的进展非常缓慢，即使是非常简明扼要的指标也比较缺乏，其中最典型的就是篇均下载频次，但即使这个指标学术界也涉及较少，缺乏深度分析。关于基于下载频次构造新的文献计量指标，有大量问题需要进一步研究：下载频次与被引频次是什么关系？下载频次的分布呈现什么规律？用多长时间跨度的下载频次来构造相关评价指标合适？如何构造下次频次相关评价指标？如何测度新指标的信息量？新指标的统计特征如何？评价中如何应用等，有必要

收稿日期：2023-10-07

基金项目：国家社科基金“学术期刊评价——指标创新与方法研究”（21FTQB016）；浙江省自然科学基金重点项目“制造业从数量型创新向质量型创新转型机制研究”（Z21G030004）

作者简介：俞立平（1967-），男，博士，教授，博导，研究方向为技术经济、科技评价领域的研究

DOI：10.13998/j.cnki.issn1002-1248.23-0757

对这些问题进行深入分析。

构造基于传播与影响力兼顾的评价指标十分重要，目前这方面指标总体比较缺乏。传播和影响力是学术期刊的两大重要标志，下载频次属于侧重传播的指标，被引频次属于侧重影响力的指标，如果能同时从这两个角度评价非常必要，当然如果设计的新指标还能一定程度上衡量学术期刊的质量、知识和信息量就更加完美了。相关研究在理论上可以丰富文献计量学，提供了一个新的评价指标，在实践中可直接采用该指标评价学术期刊的传播、影响力等综合表现，从而提供了一种新的技术手段。

关于论文相关的直接下载指标，目前中国公布了两个相关指标，一个是 Web 即年下载率，是指统计年度某期刊在中国知网发布的文献被当年全文下载的总次数与期刊论文总数之比。另一个是总下载量，某期刊发布在中国知网的所有论文在统计年被全文下载的总篇次。刘雪立[2]根据期刊引用半衰期和被引半衰期概念建立了期刊下载量半衰期。许新军[3]实证得出期刊下载量半衰期明显小于被引半衰期和引用半衰期，与被引半衰期之间存在显著的相关性。王超等[4]认为论文下载量分布可以反映相关学科文献的网络传播特征，提出最可及下载量代表指标，用来反映学科论文的下载量水平。

关于下载指标的特点与应用，HARY[5]认为下载指标具有和引文指标同样的识别重大科学发展的作用。苏新宁[6]指出下载行为源于关键词检索，被下载次数多的期刊说明其关键词比较规范，主题更贴近当前学者关注的问题。关于下载指标的影响因素，丁佐奇等[7]实证显示网络下载近 2 年比重较高，专栏及综述下载率高，研究性论文引用率高。DANIEL[8]提出下载和引文关联模型，指出文献下载受读者兴趣、文献可见性和成熟度等的影响。谢娟和龚凯乐[9]认为论文质量、引证时间窗、下载时间窗及下载数据源对下载与被引关系具有影响。

关于下载相关指标与其他指标的关系，从横向静态关系看，BOTTING 等[10]发现论文发表年内下载量可以预测今后 3 年后的被引情况，拟合优度为 0.450。SCHLÖGL 等[11]研究发现，图书情报学领域论文的下载次数与被引次数相关程度较高，相关系数达到 0.770。胡敏[12]认为从期刊层次考察，不同期刊的网络总下载量与总被引量为高度线性相关，网络篇均下载量与篇均被引量的线性相关性更强。赵一权等[13]研究结果显示计算机科学技术领域中，无论是在期刊层次，还是在文献层次，被引次数和下载次数都具有较强的正相关性。从时间趋势看，BRODY 等[14]的研究显示，论文的被下载次数对将来引用次数有积极影响。牛昱昕等[15]对开放存取论文的研究发现，从长期看下载频次与被引频次之间呈现正相关趋势。熊泽泉和段宇锋[16]研究认为，累积下载量与累积被引量存在线性相关性，且两者相关性随时间的增长而增强。

还有一些研究从更广泛的角度分析了下载指标与其他指标的关系，得出了一些独到的结论。ANDREW 等[17]对 *International Journal of Cardiology* 期刊下载频次和被引频次最高的前 25 篇论文进行比较，发现被引频次与下载频次并无显著关系。朱雯等[18]发现理、工、农、医类期刊下载频次与被引频次的相关性高于经济、人文、社会科学类期刊。陆伟等[19]实证发现下载频次与被引频次的相关性在不同情况有较大差异，单篇论文下相关性不强，作者下呈二次函数相关，而期刊下呈三次函数相关。

从现有的研究看，学术界较早意识的下载频次的重要性，并对该指标进行了大量的研究，也诞生了几个直接与下载频次相关的新指标。关于下载频次指标的特点、评价内容、关注点等研究比较充分，较多研究分析了下载频次与被引频次的关系，总体上多数研究认为下载频次与被引频次相关，同时学术界也注意到下载频次与被引频次关系的复杂性。总体上在以下方面有待深入研究。

（1）尽管下载频次比引用频次拥有更好的时效性，但采用多长时间跨度的下载频次来进行评价缺乏讨论，太长的时间跨度使得评价没有时效性，太短的时间跨度数据会导致数据不完全从而没有评价效果，有必要进行全面的数据分析后做出判断。

（2）在特定时间跨度内的下载频次，其与被引频

次的关系如何？或者换个角度，某年的被引频次数据主要受哪几年下载频次的影响？这种影响对构造新的评价指标有何影响？这也是需要进行深度计量分析。现有研究往往采用画图法、相关系数法、回归法，不足以得出有效的结论。

（3）从深度挖掘下载频次应用指标入手，新指标与影响力、传播水平、学术质量、知识和信息量等有什么关系？也需要进行进一步的分析。

（4）基于下载频次构造的新的评价指标，其统计学特征如何？对评价有什么影响？

2 基本数据分析

2.1 研究数据

本文以图书馆情报与文献学 CSSCI 期刊为研究对象（共有 20 种期刊），基于中国知网的引文数据库来进行研究，首先进行基础数据分析，进而为构造新的评价指标打下基础。考虑到论文下载频次与被引频次之间存在一定时间的数据滞后，因此载文量数据选择 2015 年，下载次数和被引次数为 2015—2021 年。需要说明的是，由于《情报学报》部分年度存在数据缺失，因此舍弃了该指标。

2.2 学科下载频次与被引频次历年变化

整个图书馆情报与文献学 2015 年发表论文的历年下载频次与被引频次如图 1 所示。下载频次在论文发表后第一年达到峰值，随后缓慢衰减，而被引频次在论文发表后第二年达到峰值，第三年略有下降，随后平稳衰减。

2.3 被引频次与下载频次滞后关系

大量研究认为被引频次滞后于下载频次 1~3 年，但这只是一种经验估计，较少有研究十分规范地分析这个问题。本文基于面板数据模型进行估计，其中一个重要的原因是，采用下载频次固然是被引频次的重要影响因素，但被引频次的影响因素太多了，由于数据所限，难以找到所有变量，在这种情况下，采用传统回归进行估计就是有偏的，但面板数据中的固定效应模型由于采用了差分估计法，使得对遗失重要变量不敏感，从而可以更好地估计下载频次与被引频次的关系。

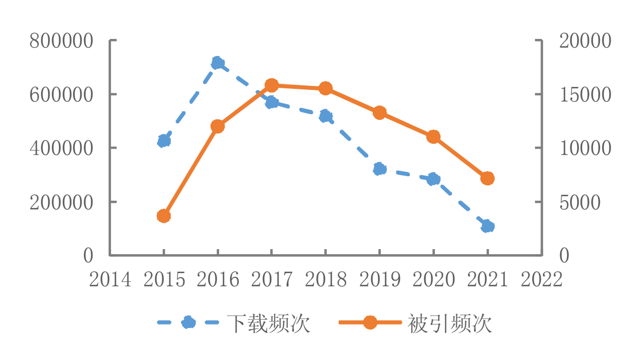


图 1　学科历年下载频次与被引频次

Fig.1 The download frequency and citation frequency of disciplines over the years

分析下载频次与被引频次的关系必须充分考虑滞后期，最后根据统计检验值是否显著来进行综合确定。由于一些期刊出版周期延长，因此本文考虑可能的滞后期为 1~5 年，基本模型如下：

$$Y = c + X(-1)\beta_1 + X(-2)\beta_2 + X(-3)\beta_3 + X(-4)\beta_4 + X(-5)\beta_5 + \mu_{it} \tag{1}$$

公式（1）中，Y 为被引频次，$X(-1)$、$X(-2)$、$X(-3)$、$X(-4)$、$X(-5)$分别表示下载频次的 1 年、2 年、3 年、4 年、5 年下载频次滞后项，β_1、β_2、β_3、β_4、β_5 分别为各自的回归系数。

面板模型的估计结果如表 1 所示，首先采用 1~5 年滞后期进行估计，先采用随机效应模型，然后进行 Hausman 检验，卡方值为 27.523，相伴概率为 0.000，拒绝随机效应的原假设，采用固定效应模型进行估计。估计结果中，4 年滞后期和 5 年滞后期回归系数为负数，3 年滞后期没有通过统计检验，明显不符合实际情况，因此选择 1~5 年滞后期是不合适的。

继续采用 1~4 年滞后期进行估计，最终采用固定效应模型，所有不同滞后期的下载频次均通过了统计检验，并且是正数。模型的拟合优度 R^2 较高，为 0.995，远远大于其他学者的预测精度，说明该模型是非常合适的，主要原因由于目前的出版周期较长，使

DOI：10.13998/j.cnki.issn1002-1248.23-0757

表 1　面板数据估计结果

Table 1　The results of panel data estimation

变量	说明	固定效应	固定效应
c	常数项	-11.427***	-8.561***
		(-6.580)	(-13.992)
ln[X(-1)]	下载次数滞后 1 年	1.173***	0.380***
		(7.690)	(3.239)
ln[X(-2)]	下载次数滞后 2 年	0.629***	0.787***
		(3.340)	(11.584)
ln[X(-3)]	下载次数滞后 3 年	0.138	0.190*
		(0.469)	(1.891)
ln[X(-4)]	下载次数滞后 4 年	-0.447**	0.101**
		(-2.543)	(2.330)
ln[X(-5)]	下载次数滞后 5 年	-0.204*	--
		(-2.022)	
hausman	hausman 检验值	27.523	10.089
p 值	相伴概率	0.000	0.039
R^2	拟合优度	0.998	0.995

注：、**、*** 分别表示在 10%、5%、1%的水平下通过统计检验

得少数论文存在较长时间的滞后。回归结果显示，2 年滞后期下载频次对被引频次的弹性系数最大，为 0.787，其次是 1 年滞后期的下载频次，弹性系数为 0.380，第三是 3 年滞后期下载频次，弹性系数为 0.190，最后为 4 年滞后期的下载频次，弹性系数为 0.101。将弹性系数转换为百分比，1 年和 2 年滞后期下载频次占 80.04%。

3　下载因子指标的构建及特征分析

3.1　下载因子确定时间窗口选取

在学术期刊评价中，评价指标的时效性非常重要。由于引文规律，使得被引频次尚未达到极大值来构建指标，即使时效性好也明显是不合理的，但如果充分考虑引文规律，导致滞后期过长，丧失评价的时效性也不合理。从图 1 可以看出，被引高峰是期刊论文发表后 2 年，影响因子的设计是非常科学的，尽管部分期刊的被引高峰是滞后 3 年。再看下载高峰，是论文发表后 1 年，如果只考虑这个因素，采用 1 年作为时间跨度是最好的，拥有比影响因子更好的时效性。

从面板数据的回归结果看，滞后 2 年的下载频次对被引频次的影响最大，滞后 1 年的下载频次对被引频次的影响次之。综合以上两个因素考虑，最终决定采用 2 年滞后期来构造下载因子指标。

3.2　下载因子的构造

基于影响因子指标的设计原理来构造下载因子 DF（Download Factor），下载因子就是期刊论文发表后两年平均每篇论文的每百次累计下载次数，用公式表示就是：

$$DF=\frac{D_t+D_{t-1}+D_{t-2}}{100P_{t-2}} \qquad (2)$$

式（2）中，t 为统计年度，D_t、D_{t-1}、D_{t-2} 分别表示统计年度、去年、前年的下载频次，P_{t-2} 为前年的可被引文献量，分母除以 100 是为了降低下载因子值的大小，更符合人们的习惯。

下载因子具有以下特点。

（1）从评价时效性看，与影响因子同步，均为期刊发表论文后 2 年。

（2）下载因子采用可被引文献量来进行计算，同样侧重学术传播和学术影响，删除了与引证计量评价无关的文献，包括科普资料、介绍、叙事抒情、摘译、摘登、转载、题要、通知、信息、资讯、导读、启事，刊首语等。

（3）为了使得下载因子具有可读性，适当降低了其数量级。

3.3　下载因子的内涵分析

下载因子的内涵如图 2 所示。首先期刊论文的知识和信息量是下载行为的原动力，它决定了下载，是学术传播的重要体现。下载频次也决定了论文的影响力，所以下载因子的内涵首先是学术传播，并且与知识信息量和学术影响力相关，表征知识信息量的典型指标就是载文量，表征学术影响力的典型指标是影响因子和 h 指数，不过影响因子通常代表一般影响力，

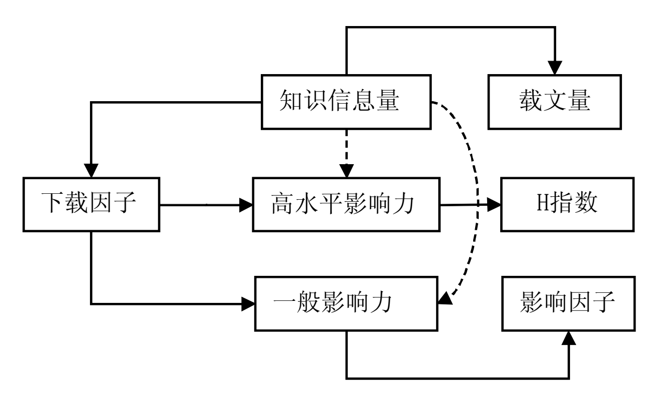


图 2 下载因子的信息含量

Fig. 2 The information content of download factors

而 h 指数代表了高水平影响力，与学术质量相关。因此基于以下模型研究下载因子的内涵组成：

$$\log(DF)=c+\alpha_1\log(H)+\alpha_2\log(IF)+\alpha_3\log(P)+\mu \quad (3)$$

公式（3）中，H 为 h 指数，IF 为影响因子，P 为载文量，μ 为随机误差项，α_1、α_2、α_3 为回归系数。

需要说明的是，分析下载因子的内涵必须在同一时间轴维度下，下载因子涉及数据是统计年度前 3 年，影响因子也是如此，因此计算 h 指数时也必须根据统计年度的前 3 年被引频次进行计算，至于载文量，是统计年度前第三年的数据。

4 下载因子的计算及实证结果

4.1 下载因子的计算结果

下载因子的计算结果如表 2 所示，排在前面的期刊包括《图书情报工作》《情报杂志》《情报科学》《图书馆学研究》《情报理论与实践》等，尽管下载因子是平均下载量指标，但下载因子较高的期刊中，仍然是载文量较大的期刊，这充分说明期刊的知识信息量对下载因子的影响较大。《中国图书馆学报》尽管载文量较低，仅有 72 篇，但下载因子仍然排在第七位，这个成绩已经十分突出。

表 2 下载因子及其他相关指标

Table 2 Download factors and other related indicators

期刊名称	下载因子	排序	2 年 h 指数	影响因子	载文量/篇
图书情报工作	19.552	1	49	5.187	765
情报杂志	14.049	2	42	5.782	463
情报科学	9.551	3	30	4.821	357
图书馆学研究	8.791	4	29	4.470	421
情报理论与实践	8.706	5	32	5.163	356
现代情报	7.740	6	31	4.382	408
中国图书馆学报	7.260	7	27	15.556	72
图书馆杂志	5.860	8	26	4.242	265
图书馆论坛	5.602	9	26	4.117	266
图书馆建设	5.052	10	24	3.174	311
图书与情报	4.541	11	26	5.534	148
大学图书馆学报	4.252	12	22	4.736	140
档案学通讯	3.695	13	18	3.418	153
图书情报知识	3.562	14	23	6.875	104
档案学研究	3.302	15	22	4.123	163
情报资料工作	3.203	16	23	3.565	170
国家图书馆学刊	3.090	17	20	4.523	153
数据分析与知识发现	2.769	18	23	2.887	230
信息资源管理学报	0.982	19	13	3.328	67

DOI：10.13998/j.cnki.issn1002-1248.23-0757

4.2 下载因子与其他指标的相关关系

考虑到影响因子、h 指数、载文量之间相关，可能存在多重共线性问题，传统的回归并不合适，因此采用岭回归来进行回归，该方法可以有效降低多重共线性问题。当标准系数之和为 0.4 时，回归结果比较稳定，因此取此时的结果作为回归结果。

$$\log(DF)=c+0.494\log(H)+0.4245\log(IF)+0.259\log(P)$$
$$R^2=0.940 \quad (4)$$

从岭回归结果看，下载因子中，对其影响最大的是 h 指数，弹性系数为 0.494，其次是载文量，弹性系数为 0.259，影响因子与其相当，弹性系数为 0.245，模型的拟合优度较高，为 0.940。换句话说，尽管下载因子表面看是一个单一指标，但其内涵信息量非常丰富，既代表了期刊的传播水平，也代表了期刊的影响力、学术质量和知识信息量，本质上是一个具有多种信息的评价指标。

4.3 下载因子的统计学性质

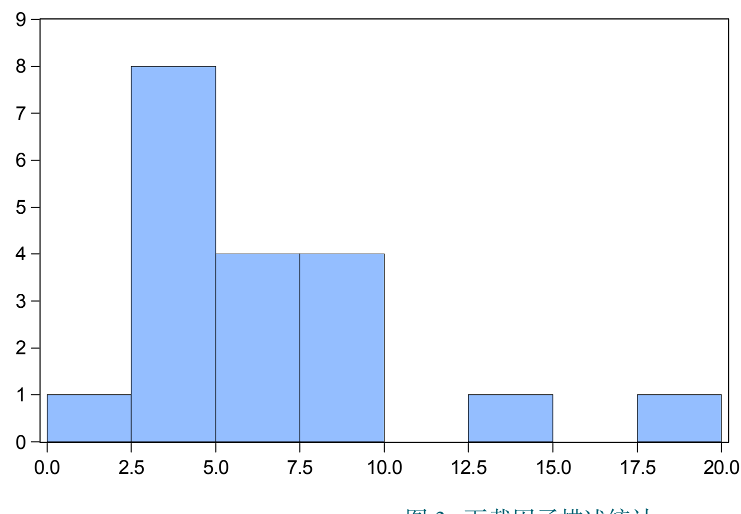
下载因子的描述统计如图 3 所示。其均值为 6.398，标准差为 4.438，拥有较好的区分度，Jarque-Bera 正态分布检验值为 12.072，相伴概率为 0.002，拒绝正态分布的原假设，其并不服从正态分布，这和影响因子、总被引频次等许多引文指标一样，也不服从正态分布。

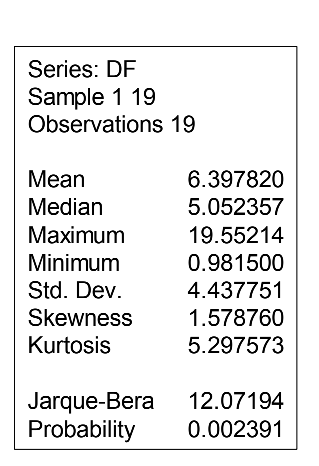
5 结论与讨论

（1）滞后 1 年和 2 年下载频次决定了被引频次的 80%。本文基于图书馆情报与文献学 CSSCI 期刊的数据研究表明，滞后 1 年和滞后 2 年的下载频次决定了被引频次的 80%，两者具有很高的拟合优度。以往学者采用相关系数和普通回归得出的主要结论是下载频次与被引频次中低度相关，主要是研究方法选择问题。本文开创性地采用面板数据模型，同时采用当期及滞后各期综合评估下载频次对被引频次的影响，从而极大地提高了预测精度。

（2）下载因子可以较好测度期刊的知识信息量、传播水平、影响力和学术质量。下载因子指标的时间轴与影响因子同步，均为期刊论文发表后 2 年，侧重学术传播水平的评估。实证研究结果表明，下载因子与表征期刊质量影响力的主要指标 h 指数相关度最高，并且与影响因子和载文量具有较高的相关性，拥有较好的统计学指标性质，是一个内涵丰富的期刊评价指标。

（3）下载因子指标有待更多学科和数据的检验。本文基于图书馆情报与文献学 19 种 CSSCI 期刊研究得出的结论，至于其他学科下载频次与被引频次的关系以及下载因子的构造问题，需要结合最新数据开展进





Series: DF	
Sample 1 19	
Observations 19	
Mean	6.397820
Median	5.052357
Maximum	19.55214
Minimum	0.981500
Std. Dev.	4.437751
Skewness	1.578760
Kurtosis	5.297573
Jarque-Bera	12.07194
Probability	0.002391

图 3 下载因子描述统计

Fig.3 The description of download factor

一步研究。

参考文献：

[1] GARFIELD E. How can impact factors be improved?[J]. BMJ, 1996, 313(7054): 411-413.

[2] 刘雪立. 科技期刊下载量半衰期的建立及其文献计量学意义[J]. 中国科技期刊研究, 2012, 23(4): 561-564.
LIU X L. Establishment of download half -life of sci -tech periodicals and its bibliometrics significance[J]. Chinese journal of scientific and technical periodicals, 2012, 23(4): 561-564.

[3] 许新军. 基于下载量的期刊半衰期实证研究[J]. 情报杂志, 2014, 33(6): 117-121.
XU X J. Empirical research on half- life period of journals based on downloads[J]. Journal of intelligence, 2014, 33(6): 117-121.

[4] 王超, 李书宁, 李晓娟. 期刊论文下载分布特征及其机制研究[J]. 情报科学, 2016, 34(12): 59-63.
WANG C, LI S N, LI X J. Research on the frequency distribution of journal paper download and its formation mechanics[J]. Information science, 2016, 34(12): 59-63.

[5] SHARMA H P. Download counts - An early indicator for monitoring progress of science[J]. Current science, 2007, 92(10): 1323-1323.

[6] 苏新宁. 构建人文社会科学学术期刊评价体系[J]. 东岳论丛, 2008, 29(1): 35-42.
SU X N. Constructing the evaluation system of academic journals of humanities and social sciences[J]. Dongyue tribune, 2008, 29(1): 35-42.

[7] 丁佐奇, 郑晓南, 吴晓明. 科技论文被引频次与下载频次的相关性分析[J]. 中国科技期刊研究, 2010, 21(4): 467-470.
DING Z Q, ZHENG X N, WU X M. Correlation analysis between citation frequency and download frequency of scientific papers[J]. Chinese journal of scientific and technical periodicals, 2010, 21(4): 467-470.

[8] O'LEARY D E. The relationship between citations and number of downloads in Decision Support Systems[J]. Decision support systems, 2008, 45(4): 972-980.

[9] 谢娟, 龚凯乐, 成颖, 等. 论文下载量与被引量相关关系的元分析[J]. 情报学报, 2017, 36(12): 1255-1269.
XIE J, GONG K L, CHENG Y, et al. Meta-analysis of the correlation between downloads and citations at paper level[J]. Journal of the China society for scientific and technical information, 2017, 36(12): 1255-1269.

[10] BOTTING N, DIPPER L, HILARI K. The effect of social media promotion on academic article uptake[J]. Journal of the association for information science and technology, 2017, 68(3): 795-800.

[11] SCHLÖGL C, GORRAIZ J, GUMPENBERGER C, et al. Comparison of downloads, citations and readership data for two information systems journals[J]. Scientometrics, 2014, 101(2): 1113-1128.

[12] 胡敏. 期刊论文网络下载规律及与引文指标相关性研究[J]. 情报杂志, 2012, 31(4): 14-18.
HU M. The law of journal papers web download and correlation of the citation index[J]. Journal of intelligence, 2012, 31(4): 14-18.

[13] 赵一权, 王振民, 熊文炳, 等. 科学论文的下载与引用关系研究: 以 ACM 数字图书馆为例[J]. 中国科技期刊研究, 2014, 25(6): 818-823.
ZHAO Y Q, WANG Z M, XIONG W B, et al. Research on the relationship between download and citation of scientific papers: Taking ACM digital library as an example[J]. Chinese journal of scientific and technical periodicals, 2014, 25(6): 818-823.

[14] BRODY T, HARNAD S, CARR L. Earlier Web usage statistics as predictors of later citation impact: Research Articles[J]. Journal of the American society for information science and technology, 2006, 57(8): 1060-1072.

[15] 牛昱昕, 宗乾进, 袁勤俭. 开放存取论文下载与引用情况计量研究[J]. 中国图书馆学报, 2012, 38(4): 119-127.
NIU Y X, ZONG Q J, YUAN Q J. A bibliometic study on downloading and citation of open access papers[J]. Journal of library science in China, 2012, 38(4): 119-127.

[16] 熊泽泉, 段宇锋. 论文早期下载量可否预测后期被引量? ——以图书情报领域期刊为例[J]. 图书情报知识, 2018(4): 32-42.
XIONG Z Q, DUAN Y F. Can downloads predict subsequent citations: A case study on journals of library and information science[J]. Documentation, information & knowledge, 2018(4): 32-42.

[17] COATS A J S. The top papers by download and citations from the International Journal of Cardiology in 2007[J]. International journal

of cardiology, 2008, 131(1): e1-e3.

[18] 朱雯, 陈荣, 刘颖. 期刊下载频次和被引频次的相关性研究——复合 H 指数视角[J]. 数字图书馆论坛, 2018(10): 25-31.
ZHU W, CHEN R, LIU Y. Relationship between citations and the number of downloads of journals: Based on compound H-index[J]. Digital library forum, 2018(10): 25-31.

[19] 陆伟, 钱坤, 唐祥彬. 文献下载频次与被引频次的相关性研究——以图书情报领域为例[J]. 情报科学, 2016, 34(1): 3-8.
LU W, QIAN K, TANG X B. Correlation analysis between document citation frequency and download frequency - In the field of library & information science[J]. Information science, 2016, 34(1): 3-8.

Journal Download Factor: A Composite Indicator of Dissemination, Impact, Knowledge and Information

YU Liping[1,2]
(1. School of Statistics and Mathematics, Zhejiang Gongshang University, Hangzhou 310018; 2. Collaborative Innovation Center of Statistical Data Engineering, Technology & Application, Zhejiang Gongshang University, Hangzhou 310018)

Abstract: [Purpose/Significance] The birth of the Internet has brought revolutionary impact on bibliometrics, giving rise to a number of online download indicators for academic literature. The most representative basic indicator among them is the download frequency, but it also includes the annual download rate, the total download volume, the download half-life, and the Google Scholar Index. The proposal of these indicators provides a new method and means of measuring scholarly dissemination and impact, which is a significant development of traditional bibliometrics and an important component of alternative metrics. Given the lack of indicators that comprehensively characterize the dissemination, impact, knowledge and information volume of academic journals, this paper proposes the download factor indicator to address this problem. [Method/Process] First, according to the changes of download frequency and citation frequency over the years, based on the citation data of CSSCI journals of library information and bibliology on CNKI, a panel data model was used to establish a prediction model of download frequency and citation frequency, and the optimal lag period for designing the download factor was determined. The indicator of download factor was proposed, that is, the average number of downloads per hundred times of each paper after 2 years of publication. This paper further used ridge regression to analyze the relationship between the download factor and the impact factor, h-index, and the number of articles. [Results/Conclusions] The download frequency with a lag of 1 year and 2 years determines 80% of the citation frequency. This article innovatively adopts a panel data model and comprehensively evaluates the impact of download frequency on citation frequency in both current and lagged periods, thereby greatly improving the prediction accuracy. The download factor can better measure the knowledge information volume, dissemination level, influence and academic quality of the journal. The timeline for downloading factor indicators is synchronized with the influencing factors, both within 2 years after the publication of journal articles, focusing on the evaluation of academic

communication level. The download factor has the highest correlation with the main indicator of the impact of journal quality, the h-index, and has a high correlation with the impact factor and publication volume. It has good statistical indicator properties and is a comprehensive indicator for evaluating journals; the download factor index needs to be more inspection of application in disciplines and use of data. This article is based on the conclusions drawn from the research of 19 CSSCI journals in library and information science literature. The relationship between download frequency and citation frequency in other disciplines, as well as the construction of download factors, require further research in conjunction with the latest data.

Keywords: download factor; level of dissemination; lag period; evaluation indicator